Semi-Synthetic Data Augmentation of Scanned Historical Documents

Romain Karpinski, Abdel Belaïd

Semi-Synthetic Data Augmentation of Scanned Historical Documents

Romain Karpinski
Télécom Nancy
54500 Vandœuvre, France
romain.karpinski@yahoo.fr

Abdel Belaïd
Université de Lorraine - LORIA
Campus scientifique
54500 Vandœuvre, France
abdel.belaid@loria.fr

***Abstract*—This paper proposes a fully automatic new method for generating semi-synthetic images of historical documents to increase the number of training samples in small datasets. This method extracts and mixes background only images (BOI) with text only images (TOI) issued from two different sources to create semi-synthetic images. The TOIs are extracted with the help of a binary mask obtained by binarizing the image. The BOIs are reconstructed from the original image by replacing TOI pixels using an inpainting method. Finally, a TOI can be efficiently integrated in a BOI using the gradient domain, thus creating a new semi-synthetic image. The idea behind this technique is to automatically obtain documents close to real ones with different backgrounds to highlight the content. Experiments are conducted on the public HisDB dataset which contains few labeled images. We show that the proposed method improves the performance results of a semantic segmentation and baseline extraction task.**

***Keywords*-Synthetic image generation; BOI; TOI; Inpainting method**

I. Introduction

Nowadays, deep learning methods assert themselves as robust methods of image processing. However, these techniques often require a lot of training data to be precise and reliable. Regarding historical documents, the amount of training data is often low because it requires the images to be analyzed and labeled by a domain expert. This labeling is time consuming therefore costly at all points. Analyzing historical documents is a challenging task since they contain a lot of difficulties. Among those difficulties, we can find heterogeneous layouts and degradations such as stains, ink drop, ink fading or missing regions. Regarding the layout, the arrangement of the lines in images makes their processing difficult.

The method employed can be summarized in three steps: First, the TOI is extracted using the binarized image. Second, all pixels composing the TOI are removed from the original image before applying an inpainting technique. As widely described in [1], inpainting is a technique inspired from art restorer. It aims to restore missing or degraded parts of an image. It can be used to recover missing regions or remove undesired objects in an image. Finally, a new semi-synthetic image is then obtained by mixing in the gradient domain TOI and BOI from different sources.

In this paper, foreground pixels are referring to ink pixels that were added to the original blank page. A TOI is a collection of foreground pixels. As opposed to foreground pixels, background pixels refers to pixels that were present in the original blank page. Therefore, a BOI is recomposed from all non foreground pixels.

The paper is organized as follows. Section II presents methods related to data augmentation and degradation models and their differences with our method. Section III describes in details the proposed method. The experiments performed with their results are described in section IV. Finally, a conclusion in section V provides a synthesis of the work and gives future trends.

II. Related work

The literature contains few works related to data augmentation, especially for historical documents. Kanungo *et al.* [3] proposed a global and local degradation model. Their model is based on real perspective distortions appearing during the scanning process. They take into account physical deformations at a global level. Then, a morphological model is used for local distortions. It changes the pixel values depending on two conditional probabilities: the probability that a foreground pixel becomes a background pixel and vice-et-versa.

To generate synthetic handwritten text lines, Varga *et al.* [4] use cosine waves and apply them to text lines images in different ways to obtain synthetic data. The results obtained showed that in the majority of the experiments, an improvement of the recognition rate was observed.

Kieu *et al.* [5] use several degradation models to generate semi-synthetic historical documents. The most common degradation models present in historical documents are employed. The ground truth is generated along the image. The method works in three steps. First, real objects are extracted from document images such as characters, background images, figures, etc. Then, degradation models (opacity, curvature, character pixel modification) are used to add noise to source images. Finally, a end-user defines the parameters for the synthetic data to be generated.

Fischer *et al.* [6] propose a method to generate training samples for historical handwriting recognition. Three degradation models are applied on binary images: Kanungo [3], character degradation from [5] and geometric distortion from

the evaluation of [7]. The results show that the best error reduction is obtained by combining the three degradation models. The error reduction of character recognition is reaching 16.53% on Saint Gall and 20.05% on Parzival dataset.

In Kieu *et al.* [8], 3D meshes are first extracted from real documents and then a set of 3D shapes is set up. These 3D shapes are used on 2D documents to generate 2D deformed document images. One disadvantage of this method is the acquisition of the 3D shapes set which requires a 3D scanner.

Seuret *et al.* [9] propose to integrate real degradation patches in the original images. The method focuses on stains integration which can be summarized in two steps: First, noise patches containing stains are manually extracted from historical document images. Then, stains are pasted onto images in the gradient domain.

We were inspired by the works of [5] and [9] by combining different image processing techniques such as object detection, inpainting and image edition in the gradient domain.

Recently, Capobianco *et al.* [10] aim to generate synthetic documents similar to existing ones. They employ a binarization technique to extract the text lines. Then, the background image is obtained by replacing text line pixels by the mean value of a window of size WxW around each pixel. The image structure is then defined by an XML file which allows to create variable structured images. To reconstruct the final synthetic images, they employ cursive fonts and dictionnaries. However, in some cases, cursive fonts may not be available and the image structure can be hard to capture when there are a lot of variations in the dataset. Similarly, Journet *et al.* [11] generate synthetic documents by extracting layout and characters using the Tesseract OCR. Then, the image is inpainted to recover the background image. The synthetic images are built according to the extracted characters, backgrounds and layouts. While this method allows precise reconstructions and combinations, when dealing with historical documents, it can be hard to segment the characters due to the cursive nature of handwriting. Both methods require the intervention of an user or a ground truth while our fully automatic method does not.

Our contribution in this work is the generation of synthetic historical documents similar to existing documents, without any user intervention, and no ground truth required to generate the images.

III. Proposed approach

As said before, the data augmentation in our case, is performed in three steps: 1) TOI Extraction, 2) BOI Extraction and 3) TOIs and BOIs Mixture. We will describe them in detail in the rest of the paper.

A. TOI Extraction

In the case of historical documents, it is desired to identify the foreground pixels of the image to isolate the text. The TOI is represented as a binary mask by binarizing the original image. While the binarization is not perfect, it can be considered as an approximation of the TOI. In this case, the quality of the TOI will rely on the binarization quality.

A crucial objective is to obtain a high recall of foreground pixels without getting too much of background pixels. In fact, a high recall is more important than a high precision because remaining foreground pixels will be present in the final image. These foreground pixels will be used during the reconstruction of the background as background pixels, which will introduce errors. For this reason, most of the time, a morphological dilation with a small structuring element is applied to the binary mask to ensure that the all foreground pixels are taken into account.

We visually compared the binarization quality with a binary mask obtained by filling the text regions represented as a polygon. The Figure 1 shows three images: (a) is the original image, (b) and (c) are respectively the binary mask obtained from the ground truth and from the binarization.

The binarization is performed using an adaptive thresholding by a sliding window of size s. The threshold is performed as the weighted mean of neighborhood values where the weights derived from a Gaussian window. A quantity c (shifting constant) is subtracted from the computed weighted mean. The parameters used for binarizing all images are $s = 31$ and $c = 21$, were empirically chosen.

(a) (b) (c)

Figure 1: Example of binary masks obtained with the two methods. (a) The original image. (b) The mask obtained from the ground truth. (c) The mask obtained from the binarization.

We can observe that the TOI obtained from a binarization is including more background pixels than the TOI obtained from the ground truth. In addition, not all foreground pixels are obtained by the binarization. This behavior was expected since the ground truth is much more precise than the binarization which considers the image as a whole while the ground truth focuses on specific areas. Once the TOI is

computed, it is used to obtain the BOI as described in the next subsection.

B. BOI Extraction

The extraction of background images is done by using an inpainting method designed to remove objects or recover missing or degraded portions of an image.

Inpainting techniques can be classified into two categories.

The first one is exemplar based where for a given region to inpaint, one looks for similar regions in available parts of the image.

The second one is diffusion based where the authors use pixels surrounding the region to inpaint to propagate the information successively.

Inpainting methods receive as inputs the image and a binary mask of the regions, to inpaint. We use the inpainting method of Telea [12] which is implemented in the computer vision library Opencv [13]. This method is diffusion based and considers the boundaries of regions to inpaint. For each point p on the boundaries, a small area $B_\epsilon(p)$ of size ϵ composed of q points is defined. For small ϵ, the first order approximation $I_q(p)$ of point p, relatively to each q point, is defined by the Equation 1.

$$I_q(p) = I(q) + \bigtriangledown I(q)(p - q) \tag{1}$$

where $I(q)$ is the image and $\bigtriangledown I(q)$ is the gradient value of point q. Each point p is inpainted using a weighted average of the first order approximation for each $q \in B_\epsilon$ (Equation 2).

$$I(p) = \frac{\sum\limits_{q \in B_\epsilon(p)} w(p,q) I_q(p)}{\sum\limits_{q \in B_\epsilon(p)} w(p,q)} \tag{2}$$

The weighted function $w(p,q)$ is defined as a product of three characteristics as shown in the Equation 3.

$$w(p,q) = dir(p,q) \times dst(p,q) \times lev(p,q) \tag{3}$$

$dir(p,q)$ is directional which increases the contribution of pixels close to the normal direction. $dst(p,q)$ is the geometric distance which decreases the impact of pixels q farther from p. $lev(p,q)$ is the level set distance which guarantees that pixel close to the boundary contributes more.

Inpainting points must be done in increasing distance order to the boundaries to mimic the way manual inpainting is done. Therefore, the fast marching method, which ensures that the point inpainted is the closest one to the known image, is employed. To inpaint the whole image, the boundary must advance one step toward its center once all its points have been inpainted. This step is repeated until there are no more boundaries to inpaint.

In the context of image documents, we seek to remove all foreground pixels and retrieve missing background pixels. Here, the image to inpaint is the original image and the binary mask is the one computed for the TOI. The Figure 2 shows BOIs using the ground truth and the binarization. BOI obtained from the ground truth TOI contains less noise because of the precision given by the ground truth. Regarding the BOI issued from the binarization, it contains remaining text pixels which are introducing noise in the BOI.

(a) (b) (c)

Figure 2: Examples of BOIs. (a) The original image. (b) The BOI obtained from ground truth. (c) The BOI obtained from binarization.

The quality of the BOI relies on two separate aspects: TOI quality and inpainting performances. The TOI quality have a huge impact since it may miss foreground pixels or include too much background pixels. For the latter, the amount of information on the background may not be sufficient to fill properly the area of the TOI. The inpainting technique used is highly responsible for the visual quality of the BOI. The method decides how to fill the regions from the TOI in the original image.

C. TOIs and BOIs Mixture

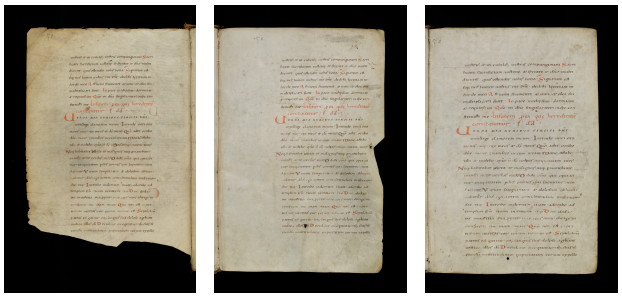
A generated image, which is semi-synthetic, is obtained by using the mixed seamless cloning as described by Perez *et al.* [14]. The authors propose several tools for editing images by using generic interpolation based on solving Poisson equations. In short, they offer an efficient solution to obtain an image from its gradient field by guiding the interpolation of the area to be changed. The mixed seamless cloning operates in the gradient domain.

Let I_a and I_b be two images where I_a is the source image and I_b the destination image with their gradient field respectively $\bigtriangledown_a$ $\bigtriangledown_b$. Let Ω be the area to insert the image and v the final gradient field. The gradients are mixed by following the equation 4.

$$\forall x \in \Omega, v(x) = \begin{cases} \bigtriangledown_b, & \text{if } \mid \bigtriangledown_b \mid > \mid \bigtriangledown_a \mid \\ \bigtriangledown_a, & \text{otherwise} \end{cases} \tag{4}$$

When combining two images, the algorithm needs to be provided with the position of one relatively to the other one. The strategy used to determine this position is to use Minimum Bounding Rectangle (MBR) of the TOIs from

both images. This means that we consider the extracted foreground pixels as foreground pixels, therefore the MBR represents the text area. To insert a given TOI in a BOI, we first compute the transformation required to transform the MBR_{TOI} into MBR_{BOI} and apply it to the TOI, i.e. we fit the new text in the area of the original text. Then, the TOI is inserted in the BOI using the mixed seamless cloning. This translation and scaling is necessary if we want to insert the text objects in the areas provided for this purpose in the layout. Ground truth annotations for the synthetic documents can be obtained by applying the same transformation on each region. In Figure 3b and 3c are two images generated with the foreground pixels of Figure 3a. One can observe that background images are different (backgrounds in Figure 3a and Figure 3b are cropped) and therefore the foreground pixels that will be added to them must be carefully placed to fit the original text location.



(a) (b) (c)

Figure 3: Example of generated images. (a) The original image. (b-c) Generated from other BOIs and the TOI of (a).

It is easy to obtain the ground truth for the generated data and it does not rely on the annotations structures. All modified objects require that the same transformation, as for inserting the TOI in the BOI, be applied to them.

Given n training images, we can augment the set by mixing every BOI with every TOI which gives at maximum n^2 augmented images. Among those n^2 images, there are n images which are the reconstruction of the original training images. These reconstructions introduce a small amount of noise and will give approximately the same results no matter what the quality of the foreground and the background separation is. This is due to the fact that the foreground pixels will be pasted in the image at the same position they were in the original one. Because we extract information and insert them back, the result can be considered as equivalent to the original. Augmenting the data with this protocol leads to a large amount of images even when the dataset is small. Another strategy is to use backgrounds from image that are not annotated to increase the background variability in augmented images. The resulting synthetic images have the following properties:

- Since TOIs are rescaled before being integrated into a BOI, the generated images offer scale variations of the text.
- Similarly, since the MBR_{TOI} is transformed to match the MBR_{BOI}, the images offer translation variation.
- By using different background images, the same text is represented in several different contexts thus highlighting it.

These properties implied by the generation method allow to introduce variabilities in synthetic images using existing images with limited variability.

IV. Experiments

A. Dataset

To conduct the experiments, we used the images from the HisDB dataset [2] subset "cb55". It consists of 20 training images, 10 validation images, 10 public test images and 10 private test images. The reported evaluations are performed on both the public and private test sets. The task selected for the experiments is part of the ICDAR 2017 competition on Layout Analysis for Challenging Medieval Manuscripts [15] (Task II) and aims to retrieve the baselines from the main text lines.

The system used to handle this task is divided in two parts:

- First, a neural network is employed to perform the semantic segmentation of the images using a fully convolutionnal neural network [16]. Semantic segmentation aims to produce class probabilities for each pixel. In our case, the two classes are either background pixel or baseline pixel.
- Second, a simple post processing step extracts the baselines points from the predicted images.

B. Generation protocol

Let T_{images} be the set of training images and $|T_{images}|$ their number. BOI from the training set are extracted to produce $|T_{images}|$ background images. This has been performed automatically with the use of the binarization technique. Let $T_{images}^{synthetic}$ be a set of size $|T_{images}|$ of synthetic images generated using random TOI from T_{images} and all background images. We produced 3 training sets to perform the experiments: $T_{exp}^{k} = T_{images} + k * T_{images}^{synthetic}$ with $k \in [0, 1, 3]$ the number of synthetic sets. Ground truth images are generated by drawing baselines with a thickness of 7 pixels.

C. Metrics

The metric used for the semantic segmentation is the F1 measure of pixels per class. Regarding the baseline evaluation, following the protocol of [15], we employed the evaluation toolkit as described in [17].

D. Neural network description

The neural network used is a U shaped fully convolutional neural network of depth 3, similar to [18]. A U shaped neural network is composed of an encoder and a symmetrical decoder. It owns skip connections which allow to use the encoded feature maps during the decoding steps. The first convolutionnal layer has 64 filters and this number is multiplied by two each time we go down of one level and divided by two when going up. To avoid memory issues with varying input sizes, we control the images shape by rescaling them arbitrary to the fixed size (720, 560). It receives as input, the normalized image and produces for each pixel $p_{i,j}$, the probability to belong to one of the two classes.

The training has been performed with the cross-entropy loss function and the Adam optimizer with a fixed learning rate of 1e-5. For all the experiments, the neural network has been trained for 200 epochs. Since classes are very unbalanced, we weighted the loss function according to the frequency of each class.

E. Baseline extraction

Prediction masks given by the previous step can result in over-segmented fragment of lines because of the thin nature of the baselines. The post processing step is performed by using a closing morphological operation to connect components horizontally in order to correct the over-segmentations. Then, a thinning operation is applied on the image to obtain a skeleton. Horizontal run-length are identified and there extreme points extracted. For each connected component, extracted points are used with the least minimum square method, to produce the final baseline.

F. Results

The results of the semantic segmentation can be observed in Table I. When comparing k_0 and k_1, it is clear that the augmentation method substantially improved the performances of the neural network. For the public test set, the background F1 measure increased of 0.7% and the baseline F1 measure of 17,33%. For the later, this is a relative improvement of 36.95%. Results on the private test are similar to the public one with an improvement of the performance of 14.42% (relative improvement 29.95%) for the baseline class. Now, considering the k_1 and k_3 sets, no further improvement has been noticed. This can be due to the fact that synthetic documents does not introduce more variability or that we reached the neural network capacities to predict baseline pixels.

The results of the baseline extraction algorithm are presented in Table II. The row labeled as *Ground truth* is here to reflect the quality of the baseline extraction technique. There is only one image where it fails partially due to the smoothing of the extracted baselines which makes them too far from the ground truth ones thus considering them as errors. This shows the theoretical maximum performances when having the perfect semantic segmentation masks. We provided to the table the results of the competition as given in [15] for comparison. We can clearly see that the data augmentation method is enabling the tested method to reach performances close to the state of the art. This demonstrates that our fully automatic synthetic data augmentation technique is sufficient to improve a system. The gain of performance is approximately equivalent to the semantic segmentation performance. The use of the data augmentation allowed to increase the performance from 81.75% to 96.59% for the public set and from 87.95% to 98.96% for the private set.

	Public		Private	
	F1 - background	F1 - baseline	F1 - background	F1 - baseline
k=0	98.80	46.9	98.76	48.14
k=1	**99.50**	64.23	**99.43**	62.56
k=3	99.45	**64.26**	99.42	**62.68**

Table I: Results of the semantic segmentation task for each experimental set.

	Public			Private		
	Recall	Precision	F1	Recall	Precision	F1
Ground truth	*98.79*	*99.33*	*99.06*	*100*	*100*	*100*
k=0	97.91	70.17	81.75	99.46	78.83	87.95
k=1	98.25	93.11	95.61	**98.7**	**96.24**	**97.45**
k=3	**98.36**	**94.89**	**96.59**	96.54	93.64	95.07
System-8	-	-	-	-	-	**98.96**
System-2	-	-	-	-	-	95.97
System-7	-	-	-	-	-	95.34

Table II: Results of the baseline extraction tasks on cb55.

Qualitative results of generated images are shown in Figure 4

V. Conclusion

In this paper, we have introduced a new generic method for generating automatically semi-synthetic data from existing images. The foreground pixels are identified and separated from the original image. Then, the BOI is retrieved using an inpainting method and the extracted foreground pixels. Finally, the semi-synthetic image is obtained by integrating TOI from one source image in the BOI of another one. This technique is simple and can easily be implemented. Since generated images contain most of existing background and foreground pixels, they look realistic regarding the other documents of the dataset and can improve the performance of a semantic segmentation task. We plan to extend the data augmentation method by studying the effect of TOI degradation before applying them to the BOI, to see how a system could be further improved. Finally, this automatic technique needs further refinement to be able to use curriculum learning by creating documents with a gradually increased difficulty.

Figure 4: Example of generated images. (a) The original image. (b) The ground truth image. (c-e) Baseline predictions for respectively k=0, k=1 and k=3.

REFERENCES

[1] S. Ravi, P. Pasupathi, S. Muthukumar, and N. Krishnan, "Image in-painting techniques-a survey and analysis," in *9th ICIIT*. IEEE, 2013, pp. 36–41.

[2] F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold, "Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts," in *15th ICFHR*. IEEE, 2016, pp. 471–476.

[3] T. Kanungo, R. M. Haralick, and I. Phillips, "Global and local document degradation models," in *ICDAR*. IEEE, 1993, pp. 730–734.

[4] T. Varga and H. Bunke, "Generation of synthetic training data for an hmm-based handwriting recognition system," in *ICDAR*. IEEE, 2003, pp. 618–622.

[5] V. C. Kieu, M. Visani, N. Journet, J.-P. Domenger, and R. Mullot, "A character degradation model for grayscale ancient document images," in *ICPR*. IEEE, 2012, pp. 685–688.

[6] A. Fischer, M. Visani, V. C. Kieu, and C. Y. Suen, "Generation of learning samples for historical handwriting recognition using image degradation," in *HIP*. ACM, 2013, pp. 73–79.

[7] J. Liang, D. DeMenthon, and D. Doermann, "Geometric rectification of camera-captured document images," *TPAMI*, vol. 30, no. 4, pp. 591–605, 2008.

[8] V. C. Kieu, N. Journet, M. Visani, R. Mullot, and J. P. Domenger, "Semi-synthetic document image generation using texture mapping on scanned 3d document shapes," in *ICDAR*. IEEE, 2013, pp. 489–493.

[9] M. Seuret, K. Chen, N. Eichenbergery, M. Liwicki, and R. Ingold, "Gradient-domain degradations for improving historical documents images layout analysis," in *ICDAR*. IEEE, 2015, pp. 1006–1010.

[10] S. Capobianco and S. Marinai, "Docemul: a toolkit to generate structured historical documents," *arXiv preprint arXiv:1710.03474*, 2017.

[11] N. Journet, M. Visani, B. Mansencal, K. Van-Cuong, and A. Billy, "Doccreator: A new software for creating synthetic ground-truthed document images," *Journal of imaging*, vol. 3, no. 4, p. 62, 2017.

[12] A. Telea, "An image inpainting technique based on the fast marching method," *Journal of graphics tools*, vol. 9, no. 1, pp. 23–34, 2004.

[13] G. Bradski, "The opencv library." *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, vol. 25, no. 11, pp. 120–123, 2000.

[14] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *ACM Transactions on graphics*, vol. 22, no. 3. ACM, 2003, pp. 313–318.

[15] F. Simistira, M. Bouillon, M. Seuret, M. Würsch, M. Alberti, R. Ingold, and M. Liwicki, "Icdar2017 competition on layout analysis for challenging medieval manuscripts," in *ICDAR*, vol. 1. IEEE, 2017, pp. 1361–1370.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[17] M. Diem, F. Kleber, S. Fiel, T. Grüning, and B. Gatos, "cbad: Icdar2017 competition on baseline detection," in *ICDAR*, vol. 1. IEEE, 2017, pp. 1355–1360.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.